Published in *SIRS'99 Conference*, 1999

Trajectory Reconstruction using Mosaic Registration

Nuno Gracias and José Santos-Victor*

Instituto Superior Técnico & Instituto de Sistemas e Robótica
Av. Rovisco Pais, Torre Norte 7.26, 1049-001 Lisboa Codex, Lisboa, Portugal

Abstract. This paper presents a set of algorithms for the creation of underwater mosaics and illustrates their use as visual maps for underwater vehicle navigation.
Firstly, we describe the automatic creation of video mosaics, which deals with the problem of image motion estimation in a robust and automatic way. The motion estimation is based on a initial matching of corresponding areas over pairs of images, followed by the use of a robust matching technique, which can cope with an high percentage of incorrect matches. Next, we deal with the issue of determining the 3D position and orientation of a vehicle from new views of a previously created mosaic. The problem of pose estimation is tackled, using the available information on the camera intrinsic parameters. This information ranges from the full knowledge to the case where they are estimated using a self-calibration technique based on the analysis of an image sequence captured under pure rotation. The performance of the 3D positioning algorithms is evaluated using images for which accurate ground-truth is available.

1 Introduction

In the last few years, computer vision has increasingly been used as a sensing modality for underwater vehicles used in tasks where accurate measures at short range are needed [11]. A considerable amount of research interest has been directed towards providing autonomy to underwater vehicles using vision, namely in self-location and motion estimation. A method for 3D motion estimation and mosaic construction was proposed by Xu *et al.* [13] and tested on a floating platform. The use of mosaics as a tool to provide visual maps for navigation as been explored by Zheng *et al.*[14], in the context of land robotics and route recognition. In their work, a visual memory of the motion of a mobile robot is created in the form of panoramic mosaics that are latter used for robot positioning. However, the visual representations are used solely for navigation purposes and the panoramic views created do not correspond to geometrically and visually correct mosaics.

The work described in this paper addresses the issues of mosaic creation and vehicle self-location using the mosaics as visual maps. The approach for automatic creation of video mosaics builds upon our previous work[7] and is based on image motion estimation in a robust and automatic way. We deal with the issue of determining the 3D position and orientation of a vehicle from new view of a previously created mosaic. The problem of pose estimation is tackled, using the available information on the camera intrinsic parameters. This information ranges from the full knowledge to the case where they are estimated using a self-calibration technique based on the analysis of an image sequence captured under pure rotation. Trajectory recovery results are presented for an image sequence for which ground-truth is available. The presented techniques are suitable for autonomous underwater vehicle (AUV) navigation near a flat oceanic floor, where a planar map is an accurate representation of the environment. A possible application scenario for these methods is in underwater archeological site exploration or in marine geological surveys, where an AUV is required to do an initial area mapping followed periodic inspections.

The paper is organized as follows. Section 2 describes some of the geometric foundations and models required for the methods used later. Mosaic creation methods are presented in Section 3, together with mosaicing results from underwater video footage. Section 4 is devoted to the registration of new views on a previously constructed mosaic and to the problem of estimating the camera pose. Finally, Section 5 summarizes and draws some conclusions on the performance and applicability of the methods.

2 Geometric background

2.1 Camera Model

In this section we will assume the reader to be familiar with the basic concepts and properties of projective geometry[3].

The most commonly used camera model in computer vision is the pinhole model under which the camera performs a linear projective mapping from the projective space $\mathbb{P}^3$ to the projective plane $\mathbb{P}^2$. This mapping can be concisely

* E-mail: {ngracias,jasv}@isr.ist.utl.pt

written as $\widetilde{\mathbf{m}} \doteq P\widetilde{\mathbf{M}}$, where $\widetilde{\mathbf{M}}$ is a 3-D point location expressed in homogeneous coordinates, $\widetilde{\mathbf{m}}$ is its projection in the retinal plane, P is the (3×4) camera projection matrix and the $\doteq$ symbol denotes the equality up to a scale factor.

An useful decomposition of the matrix P is in $P = K\,[{}^{C}_{W}R \quad {}^{C}_{W}\mathbf{t}]$, where K is the (3×3) upper-diagonal intrinsic parameter matrix, ${}^{C}_{W}R$ is the rotation matrix relating the orientation of the 3D camera and world frame, and ${}^{C}_{W}\mathbf{t}$ is the location of the world origin in camera frame coordinates. The intrinsic parameter matrix has the form

$$K = \begin{bmatrix} fk_u & fk_\theta & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where k_u and k_v are scaling factors (along u and v), and (u_0, v_0) is the location of the principal point and f is the focal length. The additional parameter k_θ gives the skew between axes. For most CCD cameras, k_θ can be considered zero on applications not relying on highly accurate calibration. For this model, the task of camera calibration can be accomplished by a simple linear least-squares minimization, using a set of six or more 3-D points of known world coordinates together with their projections on the image.

Estimation of the K matrix using the camera projection matrix The intrinsic parameter matrix K can be obtained from P by means of the QR factorization. For a 3×3 matrix A, there is a unitary 3×3 matrix Q and a upper triangular 3×3 matrix R, such that $A = QR$. For a nonsingular A, the R matrix can be chosen so as to have all positive diagonal entries. It can be shown that, in this case, the factorization is unique[9]. For the case of K matrix estimation, we are interested in the dual of the QR factorization, in the form $A = R'Q'$, where R' and Q' have the same structure of R and Q. The K matrix can be recovered from the RQ factorization of the first three columns of P. The RQ factorization for 3×3 nonsingular matrices can be computed from the QR counterpart, thus allowing the use of the number of existing implementations of the QR algorithm[10].

Let $A^T E = QR$ be the QR factorization of $A^T E$, where A is a 3×3 nonsingular matrix and

$$E = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

By means of algebraic manipulation, it can be seen that, $A = R'Q'$ where $R' = ER^T E$ is upper triangular and $Q' = EQ^T$ is unitary.

In the context of image registration and mosaicing, the collineations in $\mathbb{P}^2$ play a central role, due to the fact that two different views of the same planar scene in 3-D space are related by one. A collineation in $\mathbb{P}^2$, also commonly referred to as a planar transformation, is represented by a (3×3) matrix defined up to scale and establishing a one-to-one relation between corresponding points over two images. Thus, for a pair of image points of the same 3-D point of a planar scene with homogeneous coordinates $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{u}}'$, the collineation T_{2D} relating $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{u}}'_i$ will impose $\widetilde{\mathbf{u}}' \doteq T_{2D}\widetilde{\mathbf{u}}$. The computation of a planar transformation requires at least four pairs of corresponding points. If we have more than four correspondences, least-square minimization can then be accomplished[5, 7].

2.2 Self-calibration from a rotating camera

An alternative method can be devised for estimation of the K matrix, for the case where a sequence of images is available, taken by a camera with constant intrinsic parameters and undergoing pure rotation. This method does not require any knowledge on the scene structure, nor the rotation of the camera frame between images. Therefore, it is specially suited to applications where the camera can be rotated around its optical center, and on-line calibration is required.

The theory behind this method was presented by Hartley in [8]. For the case of stationary cameras (where no translation is allowed), the world coordinate frame can usefully be chosen to match the first camera frame. Therefore the projection matrices for cameras may be written in the form $P_i = K\,[R_i \quad \mathbf{0}]$ where $R_1 = I_3$ is the (3×3) identity matrix. The projection of a 3D world point $\widetilde{\mathbf{M}} = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T$, results in an image point $\widetilde{\mathbf{m}} \doteq K\,[R_i \quad \mathbf{0}]\begin{bmatrix} x & y & z & 1 \end{bmatrix}^T$ which is independent of the last element of $\widetilde{\mathbf{M}}$. By dropping the last element of $\widetilde{\mathbf{M}}$, the projection equations can written in the form $\widetilde{\mathbf{m}} \doteq KR_i \begin{bmatrix} x & y & z \end{bmatrix}^T$, where the reduced projection matrix $\overline{P_i} = KR_i$ performs a 2D projective mapping, similar to the planar transformations described above. Considering the case where the same 3D point $\mathbf{M} = \begin{bmatrix} x & y & z \end{bmatrix}^T$ is projected on two different images using the projection matrices $\overline{P_i} = KR_i$ and $\overline{P_j} = KR_j$, we will have $\widetilde{\mathbf{m}}_i \doteq KR_i\mathbf{M}$ and $\widetilde{\mathbf{m}}_j \doteq KR_j\mathbf{M}$, from which the following relation between $\widetilde{\mathbf{m}}_i$ and $\widetilde{\mathbf{m}}_j$ can be written,

$$\widetilde{\mathbf{m}}_i \doteq KR_iR_j^{-1}K^{-1}\widetilde{\mathbf{m}}_j = KR_{j,i}K^{-1}\widetilde{\mathbf{m}}_j$$

This equation represents a 2D homography $T_{j,i} = KR_{j,i}K^{-1}$ that maps corresponding points in two views, taken by a rotating camera. It can be computed directly from image measurements, and depends only on the intrinsic parameter matrix and on the camera rotation $R_{j,i}$ between the two images. As noted in [8], $T_{j,i}$ is only meaningfully defined up to scale, but taking into account the fact that the product $KR_{j,i}K^{-1}$ has unit determinant, the exact equality $T_{j,i} = KR_{ji}K^{-1}$ will hold if $T_{j,i}$ is scaled by an appropriate factor.

The problem remains on how to recover K and a set of rotation matrices from homographies computed from image correspondences. By using the rotation matrix property $R_{j,i} = R_{j,i}^{-T}$, we can write $T_{j,i}KK^T = KK^T T_{j,i}$. A linear system of equations can thus be constructed on the elements of the symmetrical matrix $C = KK^T$, and solved using the SVD [8].

The recovery of K can be achieved if C is positive-definite, by means of the Choleski decomposition[9] and is unique if K is assumed to have positive diagonal entries. For noise-free data, C is positive-definite by construction, and for noisy data it might not be so.

The knowledge of some of the intrinsic parameters can easily be incorporated into this self-calibration scheme with great improvement on the accuracy of the overall estimation. Namely, for the most widely used CCD video cameras and image acquisition boards, the skew between the image axes is quite often negligible. On the other hand, a rough estimate of the principal point is the central image point location, or it can be more precisely determined from radial distortion analysis[12]. For the case of zero skew and known location of the principal point, the least-squares method can still be used with straightforward modifications. Let $K = UA$ be a decomposition for a zero skew intrinsic parameter matrix, such that

$$U = \begin{bmatrix} 1 & 0 & u_0 \\ 0 & 1 & v_0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } A = \begin{bmatrix} fk_u & 0 & 0 \\ 0 & fk_v & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

From the equation $T_{j,i} = KR_{j,i}K^{-1}$ it follows that $T_{j,i} = UAR_{j,i}A^{-1}U^{-1}$ and $U^{-1}T_{j,i}U = AR_{j,i}A^{-1}$. By conducting the same algebraic manipulation as above, a similar system of equations can be constructed for the elements of the diagonal matrix $D = AA^T$.

3 Mosaic Creation

3.1 Feature selection, correlation matching and robust selection

The work presented here evolves from the analysis of point projections and their correspondence between image frames. The selection of image points to be matched is based on a simplified version of the Harris corner detector. Next, the point features are matched over the next frame, by a correlation-based matching procedure using the sum of squared differences (SSD). In order to reduce the computational burden of the matching, the search is restricted to an area whose location and size are given by the analysis of the cross-correlation between the two images.

The result of this matching process are two lists of coordinates of corresponding points. Due to the error prone nature of the correlation matching, it is likely that a number of point correspondences will not relate to the same 3-D point. Therefore, a robust selection of the matched points is required for accurate image motion estimation. For this, a random sampling algorithm was implemented, using the following minimization criterion. Let ${}^{(k)}\mathbf{u}_i$ be the location of the i^{th} feature extracted from image I_k, and matched with ${}^{(k+1)}\mathbf{u}$ on image I_{k+1}. The criterion to be minimized is the median of sum of the square distances,

$$\underset{i}{\text{med}} \left(d^2\left({}^{(k)}\mathbf{u}_i, T_{k,k+1}\ {}^{(k+1)}\mathbf{u}_i\right) + d^2\left({}^{(k+1)}\mathbf{u}_i, T_{k,k+1}^{-1}\ {}^{(k)}\mathbf{u}_i\right)\right) \tag{1}$$

where $d(\cdot,\cdot)$ stands for the image point-to-point Euclidean distance and $T_{k,k+1}$ is the planar transformation matrix relating the frames k and $k+1$.

The random sampling algorithm is a simple two-step variant of least-median-of-squares (LMedS), comprising two phases of random sampling LMedS. After the first phase, the data set is reduced by selecting the best data points in the sense of the chosen cost function. Next, the reduced data undergoes another random sampling LMedS phase. Results on the use of this algorithm for general-purpose mosaicing and fundamental matrix estimation are presented in [5] and [6].

3.2 Registration and rendering

After estimating the frame-to-frame homographies, these are cascaded to form a global registration, where all frames are mapped into a common, arbitrarily chosen, reference frame. Let $T_{Ref,1}$ be the transformation matrix relating the chosen reference frame and the first image frame. The global registration is defined by the set of transformation matrices $\{T_{Ref,k} : k = 1 \ldots N\}$, where for $2 \leq k \leq N$,

$$T_{Ref,k} = T_{Ref,1} \prod_{i=1}^{k-1} T_{i,i+1}$$

In this work, the chosen reference is computed using some points with known metric coordinates. Let us assume that a set of 4 or more points have known positions on a 2D world coordinate frame, and that those points have identifiable projections on the first image of the sequence. The homography $T_{World,1}$ relating the 2D world frame and the first image frame can, therefore, be computed using the same least-squares method as used for the inter-image homographies.

After global registration, the following step consists in merging the images. On overlapping regions there are more multiple contributions for a single point on the output image, and some method has to be established in order to determine the unique intensity value that will be used. For the mosaicing example given in this paper, the median operator was used, which is adequate for underwater sequences of the seabed, where moving fish or algae are captured. Figure 1 presents two views of a mosaic from a sequence of images captured by a surface-driven ROV, on a pipeline inspection task. In this example the perspective distortion effects are noticeable, since the image plane of the camera is distinctly not parallel to the sea floor. No information was provided about the camera motion, which is composed of translation and rotation. The left image was created using the first frame of the sequence as the reference frame. For the right image, a reference frame was chosen as to make the contour lines of the pipeline approximately parallel, yielding a top view of the floor.

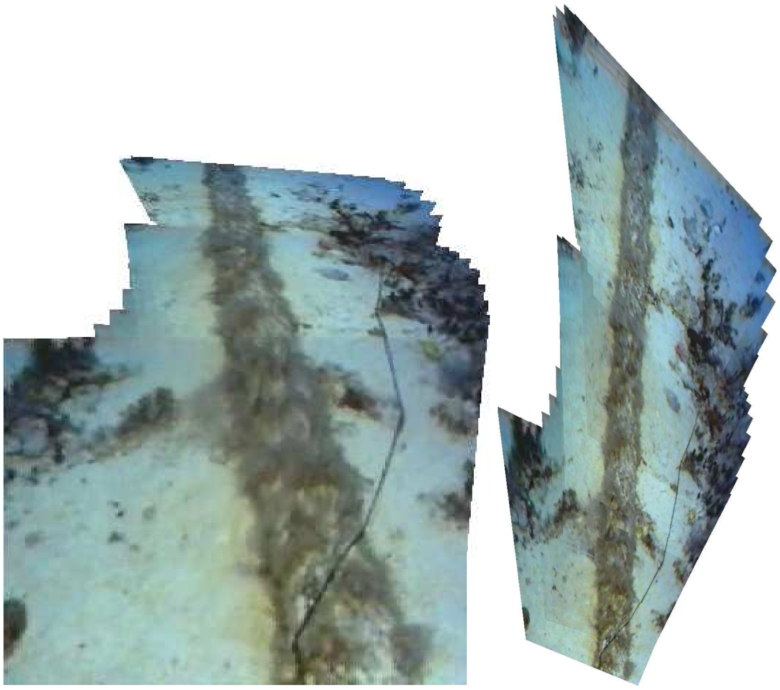

Fig. 1.: Underwater pipeline mosaic example. For the image registration, the full planar tranformation model was used. The images were registered with the use-last operator (left). A useful reference frame can be chosen in order to have a better perception of the sea floor (right).

4 Pose Estimation

For an underwater navigation application, we are interested in using video mosaicing as a tool to provide visual navigation maps. Having such maps referenced to a world coordinate system, will enable a camera-equipped autonomous vehicle (in an unknown position and orientation) to locate itself once it has found the correct mapping from the mosaic to the image frame. Some methods will be presented in this Section for the pose estimation. Trajectory recovery is illustrated, based on the assumption of constant intrinsic parameters

Over the years, the problem of camera pose estimation has been thoroughly addressed in the Computer Vision literature. For recent progress in linear methods in pose estimation refer to [1] and the references there in. The pose estimation algorithm presented in this Section decomposes of an image-to-mosaic homography matrix, in order to find the rotation matrix and displacement vector relating the camera frame to a world frame (extrinsic parameters). In this sense, it relates to the work by Ganapathy[4], where the extrinsic parameters are recovered directly from a camera projection matrix.

4.1 Algorithm for new view registration on the mosaic

We will now deal with the problem of finding the homographies relating a sequence of images to a previously constructed mosaic. In this, we explicitly take advantage of the timely order nature of the image sequence, to reduce the computational burden of finding correspondences on the mosaic. By assuming that adjacent image frames correspond to a small camera translation and rotation, we have a large image overlap. Thus, the feature matching search is restricted to a neighboring area of the location predicted by the last image-to-mosaic homography.

One of the main difficulties with this procedure is due to the fact that feature matching using correlation produces poor results in the presence of significant image warping between the image features and the mosaic. However this condition can be greatly alleviated by using an estimate of the image-to-mosaic homography, and performing feature warping before correlation.

The implemented algorithm requires an estimate of the first image-to-mosaic homography, which needs not be too accurate. For each image, the algorithm tries to find a reliable image-to-mosaic homography. Reliability is insured by the specification of a minimum acceptable number of correct matches. If it fails to find that number, the algorithm uses the homography with the previous image to compute infer the current image-to-mosaic homography. The advantage of registering each frame directly on the mosaic (as opposed to computing by sequentially cascading the homographies $T_{i-1,i}$, between previous images), is due to the fact that small estimation errors on $T_{i-1,i}$ are not on accumulated. This condition is apparent on the results in Section 4.4.

Once the image-mosaic homographies have been computed and the mosaic is referenced to a world frame, the camera pose can be estimated with respect to the world frame. Four methods have been implemented for this, which differ on the amount of information available on the intrinsic parameters of the camera. They are:

- Completely known intrinsic parameter matrix.
- Known principal point and skewing.
- Unknown intrinsic parameter matrix, but estimated using self-calibration from rotating scenes.
- Unknown intrinsic parameter matrix, estimated using self-calibration from rotating scenes with additional knowledge on the principal point and skew.

The last two methods are extentions of the first. Under these two, the self-calibration scheme described earlier is used to estimate the K matrix. Afterwards, the first method is used to recover the trajectory. In practical setups, this implies an additional manoeuvre using a pan and tilt head mounted on the vehicle, or the possibility of the vehicle rotating maintaining the camera optical center approximately at the same position.

4.2 Known intrinsic parameter matrix

Throughout the rest of this section we will consider a 3-D world frame to be such that all the points in the planar scene have null $\overrightarrow{z}$ coordinate. Let J be the (4×3) matrix obtained from the (4×4) identity matrix by suppression of its third column, and $P = K\,[{}^{C}_{W}R \quad {}^{C}_{W}\mathbf{t}]$ the projection matrix of a camera whose optical center is not contained by the plane. It can easily be seen that the relation between the planar points and their projections in the image plane is described by an homography given by $T_{image,World} \doteq PJ$. This homography can be expressed in the form

$$T_{image,World} \doteq KL \tag{2}$$

where the L matrix comprises the first two columns of ${}^{C}_{W}R$, and the vector, ${}^{C}_{W}\mathbf{t}$. The L matrix can be computed up to scale, from $L \doteq K^{-1}T_{image,World}$, since $T_{image,World}$ is also only defined up to scale.

In order to recover the ${}^{C}_{W}R$ matrix, one has first to estimate the unknown scale factor λ. This factor effects the matrix L, therefore scaling the first two columns of ${}^{C}_{W}R$. Let $\mathbf{r}_1$ and $\mathbf{r}_2$ be the first two columns of L. It is easy to see that the absolute value of λ is given by the norm of $\mathbf{r}_1$ (or $\mathbf{r}_2$), due to the fact that the columns of a rotation matrix have unit norm. In this work we have taken $\|\lambda\|$ to be the average of the two norms.

A simple way of recovering the rotation matrix ${}^{C}_{W}R$ is, firstly, to scale $\mathbf{r}_1$ and $\mathbf{r}_2$ by the two symmetric solutions for λ and, secondly, to compute their cross-product. This cross-product is, in fact, the same for both solutions. Since $T_{image,World}$ is estimated from noisy image measurements, it is to expect $\mathbf{r}_1$ and $\mathbf{r}_2$ to be non-orthogonal. In this work, the orthogonality condition was enforced by computing the two orthogonal unit vector that have the same bisectrix and stand on the same plane as $\mathbf{r}_1$ and $\mathbf{r}_2$.

Let ${}^{C}_{W}R_1$ and ${}^{C}_{W}R_2$ be the two candidates for ${}^{C}_{W}R$, corresponding respectively to the scaling by $+\|\lambda\|$ and $-\|\lambda\|$. The corresponding optical centre locations are given by

$${}^{W}_{C}\mathbf{t}_1 = -\frac{1}{\lambda}\,{}^{C}_{W}R_1 \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} \text{ and } {}^{W}_{C}\mathbf{t}_2 = \frac{1}{\lambda}\,{}^{C}_{W}R_2 \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$$

It can easily be seen that ${}^{C}_{W}R_1$ and ${}^{C}_{W}R_2$ relate by ${}^{C}_{W}R_1 = {}^{C}_{W}R_2 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, which represents a rotation of 180 degrees around the $\overrightarrow{z}$ axis. Therefore the locations of the optical centres differ by the last coordinate which is symmetric.

Both solutions for ${}^{C}_{W}R$ and ${}^{W}_{C}\mathbf{t}$ are coherent with $T_{image,World}$, and therefore valid. In the application of this work, we are only interested in the positive $\overrightarrow{z}$ axis solution for ${}^{W}_{C}\mathbf{t}$, which corresponds to the camera being above the plane of the floor.

4.3 Known principal point and skewing

An alternative method for estimating the camera pose can be devised if only the principal point location and the skewing ratio $\frac{fk_\theta}{fk_v}$ are known, instead of the full K matrix. Let us decompose K as the product of an upper triangular matrix U, with ones on its diagonal, by a diagonal matrix A, such that

$$K \doteq UA = \begin{bmatrix} 1 & \frac{fk_\theta}{fk_v} & u_0 \\ 0 & 1 & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} fk_u & 0 & 0 \\ 0 & fk_v & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since U is invertible, one can extend Equation (2) to

$$U^{-1}T_{image,World} \doteq AL \tag{3}$$

The left side of Equation (3) can be computed from image measurements. As we are interested in estimating the unknown intrinsic parameters in the A matrix, we will start by explicitly including an unknown scale factor λ, in order to remove the equality up to scale. Let $M = U^{-1}T_{image,World}$. Equation (3) can thus be written as $A(\lambda L) = M$. By considering the first two columns of this equality, where the unknown scale factor λ has been multiplied to the elements of L, the following system of equations can be written,

$$\begin{bmatrix} fk_u & 0 & 0 \\ 0 & fk_v & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda \cdot l_{11} & \lambda \cdot l_{12} \\ \lambda \cdot l_{21} & \lambda \cdot l_{22} \\ \lambda \cdot l_{31} & \lambda \cdot l_{32} \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \end{bmatrix}$$

By imposing the additional conditions of equal norm and vector orthogonality on $\begin{bmatrix} \lambda \cdot l_{11} & \lambda \cdot l_{21} & \lambda \cdot l_{31} \end{bmatrix}^T$ and $\begin{bmatrix} \lambda \cdot l_{12} & \lambda \cdot l_{22} & \lambda \cdot l_{32} \end{bmatrix}^T$, a system of equations on fk_u and fk_v can be written in the form of

$$\begin{bmatrix} m_{11} \cdot m_{12} & m_{21} \cdot m_{22} \\ m_{11}^2 - m_{12}^2 & m_{21}^2 - m_{22}^2 \end{bmatrix} \begin{bmatrix} \frac{1}{fk_u^2} \\ \frac{1}{fk_v^2} \end{bmatrix} = - \begin{bmatrix} m_{31} \cdot m_{32} \\ m_{31}^2 - m_{32}^2 \end{bmatrix}$$

After estimating fk_u and fk_v the pose can be recovered using the method described in Section 4.2.

Although this method can be used independently for each frame of the image sequence, higher accuracy for the pose recovery can be achieved by using more than one frame in the estimation of fk_u and fk_v. For the experiments conducted in this paper, we have used an iterative least-squares method[2], which is suitable for the on-line processing of image sequences.

4.4 Pose Estimation Results

Original Sequence In order to evaluate the performance of the pose estimation algorithms, accurate ground-truth is required. For this reason we have used the mosaic of Figure 1 and synthesized new views according to a specified camera matrix and trajectory. These images are then used to retrieve the camera and position parameters. The mosaic was set to cover an area of 6 by 14.5 meters. The sequence comprises 40 images of 320×240 pixels taken by a camera on a moving vehicle combining 3D motion and rotation. The camera is pointing downwards with a tilt angle of approximately 150 degree with respect to the horizontal. The used intrinsic parameters matrix K accounts for a skewless camera with the following intrinsics,

$$K = \begin{bmatrix} 480 & 0 & 160 \\ 0 & 480 & 120 \\ 0 & 0 & 1 \end{bmatrix}$$

In order to simulate the vehicle drift induced by water currents, perturbations have been added to the nominal forward motion of 0.23 meters/frame and to a nominal height above sea floor of 3 meters. The perturbations account for periodic drifts of around 0.4 meters in position and 15 degrees in orientation. For each frame, the combined movement of the camera is depicted on Figure 2, where the camera is represented by its optical axis.

The resulting synthetic sequence is fairly realistic-looking. Frames 9 and 18 are presented on Figure 3. Although it does not take into account the effects of non-uniform lighting and barrel distortion, usually present in underwater imagery, this sequence can still be a good approximation of real images taken in shallow waters under daylight illumination.

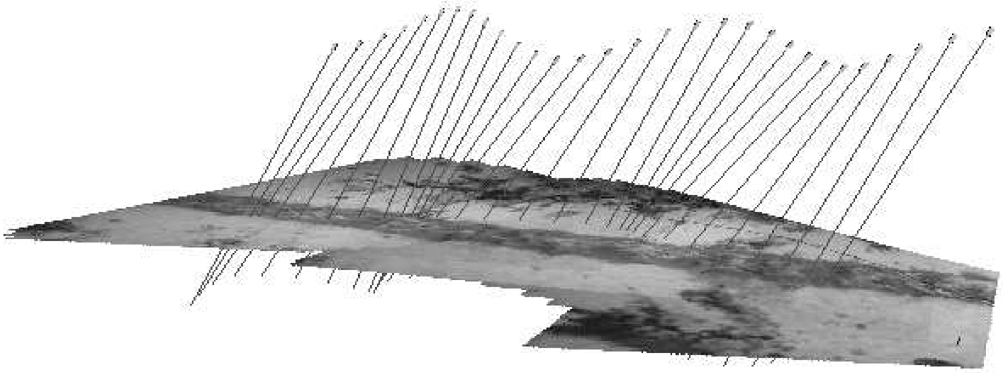

Fig. 2.: 3D view of the camera positions and corresponding optical axes used for generating the sequence with available ground-truth..

Fig. 3.: Two frames of the generated image sequence used for positioning evaluation with ground-truth.

Camera Self-calibration In order to use the self-calibration method, an additional set of 20 images was produced, in which the camera undergoes pure rotation. The optical centre remained fixed at 4 meters above the sea bottom, while the camera faced down, and rotated around the 3 axes (pan, tilt and yaw). For each axis, the angle range is ± 5 degrees. The intrinsic parameters matrix K used for this sequence was the same as the one used for the other sequence.

The estimation of the homographies between adjacent images constitutes the starting point for the self-calibration procedures. The homographies were computed using the algorithms described above for the mosaic creation. For the sequence of 20 images, 19 homographies were estimated using 6 to 70 matched pairs of points, within a 0.5 pixel distance threshold. Using the self-calibration methods for unknown intrinsics and known principal point and skew, the recovered matrices were, respectively,

$$K_{rec} = \begin{bmatrix} 499.8 & -10.1 & 159.2 \\ 0 & 482.9 & 82.8 \\ 0 & 0 & 1.00 \end{bmatrix} \text{ and } K_{recPP} = \begin{bmatrix} 484.38 & 0 & 160.00 \\ 0 & 487.33 & 120.00 \\ 0 & 0 & 1.00 \end{bmatrix}$$

Camera Trajectory Recovery The performance of the pose estimation was experimentally evaluated by testing the camera path reconstruction using the available ground-truth data. The test results presented assume constant intrinsic parameters in time and differ on the amount of intrinsic parameter information used. With decreasing required calibration information, the conducted tests were the following:

Exp1 - Trajectory recovery with known K matrix
Exp2 - Trajectory recovery with known principal point and zero skew
Exp3 - Trajectory recovery with self-calibration with known principal point and zero skew
Exp4 - Trajectory recovery with self-calibration for all intrinsic parameters

The synthetic images from the sequence containing camera translation were registered directly on the mosaic, using the algorithm described in subsection 4.1. The algorithm was run with a specified acceptable minimum of eight match pairs per homography. In each frame it was able to find between 16 and 39 pairs. For the set of 40 images, 2 homographies were computed with matched pairs from a second attempt, while the other 38 were computed at the first attempt.

For the trajectory recovery with known principal point and zero skew (Exp2), an iterative least-squares estimation method was implemented for the estimation of the two unknown intrinsic parameters. When processing a sequence of images, this method uses the current frame-to-mosaic homography, together with all the past homographies.

Statistics for the reconstruction errors are presented in Table 1. The position errors were measured by taking the Euclidean distance between the ground-truth position and the estimated position. As for the orientation, the error was measured by computing the angle between the true and estimated camera frame orientations. For each image and method, the position errors are plotted in the left side of Figure 4.

Method	Position Errors (meters)			Angular Errors (degrees)		
	Avg.	Maximum	Std.Dev.	Avg.	Maximum	Std.Dev.
Known K (Exp1)	0.031	0.159	0.031	0.610	2.932	0.525
Known PP (Exp2)	0.045	0.159	0.030	0.636	2.978	0.509
Self-Calib. with PP (Exp3)	0.061	0.163	0.025	0.690	2.675	0.463
Full Self-Calib. (Exp4)	0.258	0.366	0.036	1.678	2.754	0.421

Table 1.: Trajectory recovery results for known K matrix, known principal point, self–calibration with known principal point and full self-calibration. Average, Maximum and Standard Deviation of position and angular errors.

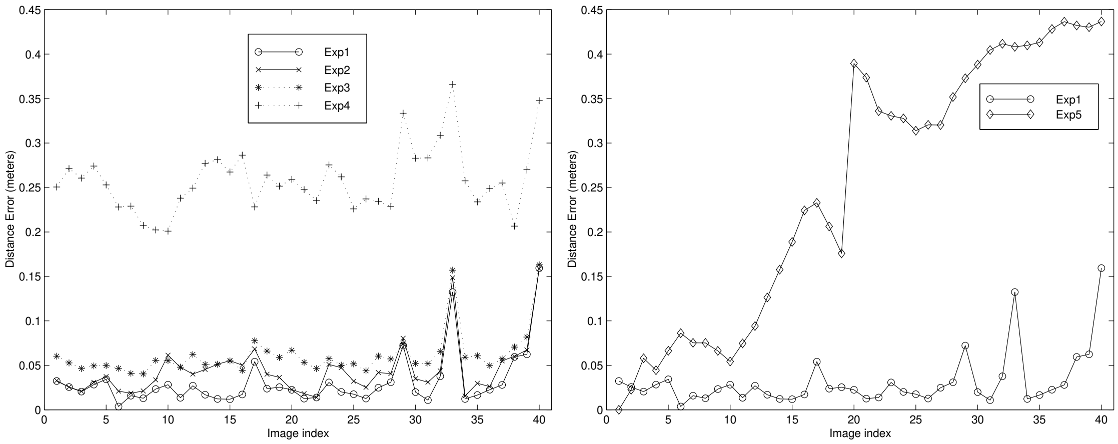

Fig. 4.: Trajectory position errors for the experiments 1 to 4 (left) and experiments 1 and 5 (right). For Exp5, the error growth is apparent, resulting from the accumulation of small errors during the inter–image homography estimation.

In these results, the lowest position and orientation errors correspond to the trajectory recovery with known K matrix (Exp1). This is not surprising, as this method uses the most prior information. For Exp2, a 50 % increase on the average error is observed. With self-calibration for known principal point and zero skew (Exp3), the average position error is approximately 2 times larger than the one for Exp1. However the worst case error is 15 cm, which can be considered small when compared to the distance to the sea floor of 3 meters (5%). The average error is 6.1 cm which accounts for slightly more than 2% of the distance to the sea floor. When using unconstrained self-calibration (Exp4), position errors of around 25 cm are observed.

An additional experiment was conducted in order to compare the following image registration schemes:

Exp1 - Image-to-mosaic homographies computed by direct mosaic registration
Exp5 - Image-to-mosaic homographies computed by cascading inter-images homographies

The first scheme refers to the use of the algorithm of Subsection 4.1 with the same setup as in (Exp1). In the second, the true camera position and orientation is used for computing the first image-to-mosaic homography $T_{M,1}$. The subsequent homographies are calculated by,

$$T_{M,i} = T_{M,1} \cdot \prod_{k=2}^{i} T_{k-1,k} \qquad i > 1$$

where $T_{k-1,k}$ are the inter-images homographies and the matrix product is computed by right-multiplying for each increment of the index k. The set of $T_{k-1,k}$ was estimated from the same sequence of images, and the number of used matched points varied from 10 to 76 pairs, with an average of 60.

Table 2 and Figure 4 (right) present, respectively, the observed position and orientation error statistics, and the plot of the positions errors for each frame. It can be seen that the second scheme produces less accurate results, due to the fact that small errors, inherent to the inter-image homography estimation, are accumulated. This phenomena is in many ways comparable to the positioning errors arising from the use of dead-reckoning during navigation.

Method	Position Errors (meters)			Angular Errors (degrees)		
	Avg.	Maximum	Std.Dev.	Avg.	Maximum	Std.Dev.
Direct Reg. (Exp1)	0.031	0.159	0.031	0.610	2.932	0.525
Inter–image Reg. (Exp5)	0.252	0.437	0.151	2.667	5.528	1.313

Table 2.: Error statistics for the direct mosaic registration scheme (same as in Exp1) and for the inter-image registration scheme. The values of the average, maximum and standard deviation are shown for the position and angular errors .

A final experiment was conducted for evaluating the quality of reconstruction using self-calibration when the principal point location is known or estimated beforehand with a different method. For the set of synthetic images described above, the self-calibration algorithm with defined principal point was used, with the location of the principal point varying from (0,0) to (320,240) in equally spaced intervals of 1 pixel. For each location, the average error on camera position and orientation was computed. Contour plots showing the results for average errors are given in Figure 5.

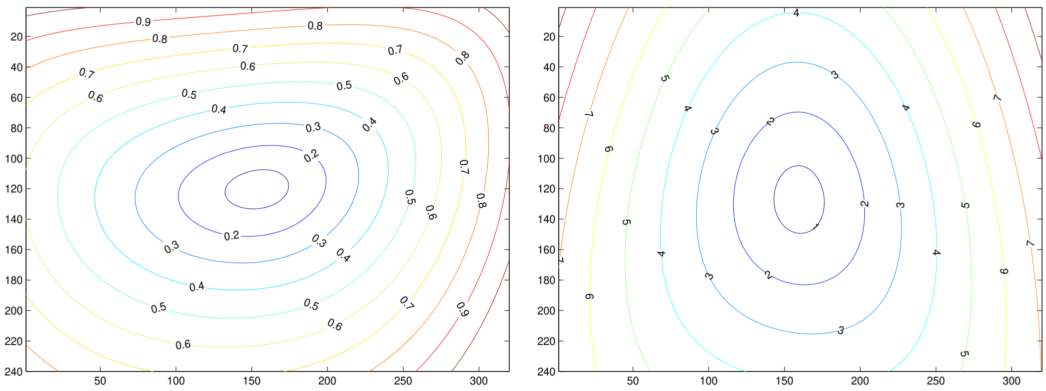

Fig. 5.: Results for the trajectory recontruction procedure using self-calibration from rotating camera with known principal point. These contour plot contains the average position error in meters, (left) and the average angular errors in degrees (right), as a function of the assumed principal point location.

The lowest errors correspond to the true principal point at (160,120), at the center of the graphs. This case reverts to the third experiment (Exp3) described above, for which the average position and orientation errors were 0.061 meters and 0.690 degrees. The method used in that experiment outperforms the unknown intrinsics self-calibration method (Exp4), for a considerably large area around the principal point.

From this, a conclusion with practical implications for the setups where self-calibration is feasible can be drawn. Even if the principal point location is not known precisely, it might be more advantageous to use a rough estimate and perform self-calibration for known principal point than to do it for unconstrained intrinsics. For the case presented here, better results on position error are achieved even if the assumed principal point location is within a 35 pixel neighborhood of correct one.

5 Conclusions

We have presented an approach for the use of underwater video mosaics as visual reference maps for vehicle localization. Key issues for the mosaicing process are the robust selection of correspondences and the use of geometric models capable of registering any view of a planar scene. Methods for pose estimation were presented, which allow the estimation of the 3D position and orientation of a vehicle from a view of a previously created mosaic. The performance was evaluated using images corresponding to known camera motion that served as the ground-truth.

An emphasis was put on using several degrees of available information on the camera intrinsic parameters, including self-calibration. The possibility of calibrating a camera *on-line* can be of practical importance for a number of visually guided tasks, specially if the camera parameters are subject to change slowly in time. This paper illustrated how relevant information for the pose estimation process can be obtained by the analysis of rotation images. These images are easier to acquire than having to resort to calibration grids. Also, we have shown that the knowledge of the principal point can easily be incorporated in the self-calibration scheme with benefits on the accuracy. This is true even if its location is not precisely known.

By automatically creating visual representations of the sea floor and using them for navigation, the methods described in this paper provide an important step towards the autonomous operation of submersibles.

Acknowledgements

The work described in this paper has been supported by the Portuguese Foundation for Science and Technology PRAXIS XXI BD/13772/97, NARVAL Esprit-LTR Proj. 30185 and INFANTE Proj. PRAXIS 2/2.1/TPAR/2042/95.

References

1. J. Batista, H. Araújo, and A. Almeida. Iterative multi–step explicit camera calibration. In *Proc. of the 6th International Conference on Computer Vision*, Bombay, India, January 1998.
2. C. Brown. Tutorial on filtering, restoration and state estimation. Technical Report 534, University of Rochester, Computer Science, June 1995.
3. O. Faugeras. *Three Dimensional Computer Vision*. MIT Press, 1993.
4. S. Ganapathy. Decomposition of transformation matrices for robot vision. In *Proc. 1st IEEE Conf. Robotics*, pages 130–139. IEEE, 1984.
5. N. Gracias. Application of robust estimation to computer vision: Video mosaics and 3–D reconstruction. Master's thesis, http:// www.isr.ist.utl.pt/labs/vislab/thesis, Lisbon, Portugal, April 1998.
6. N. Gracias and J. Santos-Victor. Robust estimation of the fundamental matrix and stereo correspondences. In *Proc. of the International Symposium on Intelligent Robotic Systems*, Stockholm, Sweden, July 1997.
7. N. Gracias and J. Santos-Victor. Automatic mosaic creation of the ocean floor. In *Proc. of the IEEE OCEANS'98*, Nice, France, September 1998.
8. R. Hartley. Self–calibration from stationary cameras. *Internation Journal of Computer Vision*, 22(1):5–23, February/March 1997.
9. R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
10. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1988.
11. J. Santos-Victor and J. Sentieiro. The role of vision for underwater vehicles. In *Proc. of the 1994 Symposium on Autonomous Underwater Vehicle Technology*, pages 28–35, Cambridge, MA, USA, July 1994.
12. R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv camera and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, 1987.
13. X. Xu and S. Negahdaripour. Vision-based motion sensing for underwater navigation and mosaicing of ocean floor images. In *Proc. of the Oceans '97 Conference*, volume 2, pages 1412–17, Halifax, Canada, October 1997.
14. J. Zheng and S. Tsuji. Panoramic representation for route recognition by a mobile robot. *International Journal of Computer Vision*, 9(1):55–76, October 1992.

This article was processed using the TEX macro package with SIRS99 style